Using Machine Learning - Find Novel Intrinsic Oncology Targets and Biology

Xiaoyuan Liu, Haoen Huang, Yujia Hong, Liuyi Cui
Department of Mathematics, Northeastern University

Abstract

Classification serves as a potent instrument in machine learning, enabling precise decision-making and predictive modeling. Utilizing labeled training data, classification algorithms discern patterns and associations within the data, subsequently assigning novel, unobserved instances to their appropriate categories.

This project endeavors to employ a range of machine learning methodologies to analyze CRISPR technology data and classify it into distinct cancer lineages. Our preliminary findings have been encouraging, prompting the team to concentrate on lymphoid cancer specifically. The ultimate objective is to attain precise learning outcomes despite the constraints imposed by limited gene data.

1 Introduction

CRISPR Technology

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology is an influential genetic tool that enables scientists to edit genes with remarkable accuracy. The process involves employing a guide RNA to target a specific DNA sequence and utilizing an enzyme called Cas9 to cleave the DNA at the targeted site. Researchers can then either introduce new DNA to replace the original sequence or allow the cell to naturally repair the cut, potentially altering the genetic code.

Classification in Machine Learning

Classification constitutes a fundamental task in machine learning, aiming to predict the class or category of a given input based on a set of features. The primary objective of classification is to develop a model capable of accurately predicting the class of new, unseen data by learning patterns from a labeled dataset.

Supervised learning is the prevalent approach to classification, wherein the model is trained on a labeled dataset containing examples of inputs and their corresponding classes. This labeled data enables the model to identify patterns and associations between features and class labels. Once trained, the model can predict the class of new, unseen data.

In the context of CRISPR technology, classification can be applied to identify essential genes. CRISPR screens can be employed to disable specific genes in a cell and observe the resulting phenotype. Classifying genes based on their phenotype allows researchers to pinpoint essential genes necessary for cell survival or growth.

Furthermore, classification models can be utilized to predict the lineages of uncharacterized cells. By training the model on a dataset comprising genes with known functions and their corresponding features, it can learn to classify new genes based on their features and anticipate their potential lineages.

2 Methods

2.1 Machine Learning Algorithms

Various algorithms are employed for classification, including logistic regression, support vector machines, and neural networks. Each algorithm possesses unique strengths and weaknesses, making them suitable for different types of datasets.

2.1.1 Classification

Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant Analysis

Logistic Regression is a widely-used machine learning algorithm for binary classification problems. It utilizes a logistic function to convert input features into probabilities of belonging to one of two classes. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) emphasize linear separations, differing in their assumptions about covariance.

Artificial Neural Networks and Support Vector Machines

Artificial Neural Networks (ANNs) are a category of models that learn from data by constructing interconnected layers of artificial neurons. In contrast, Support Vector Machines (SVMs) are algorithms that identify the hyperplane optimally separating data into distinct classes. Both ANNs and SVMs can manage high-dimensional data and learn intricate patterns within the data.

2.1.2 Regression

Random Forest

Random Forest is a widely-adopted ensemble learning algorithm utilized for both classification and regression tasks. It functions by constructing multiple decision trees during training and outputs either the mode of the classes (classification) or the mean prediction (regression) of individual trees.

2.2 DepMap Public 22Q4 Primary Files

The DepMap Public 22Q4 Primary Files dataset is supplied by the Broad Institute's DepMap project, which seeks to delineate genetic dependencies in cancer. This dataset encompasses genomic and drug sensitivity data for over 1,000 cancer cell lines originating from various tissues, such as breast, lung, colon, and more. The dataset is available in multiple formats, including compressed TSV files and SQLite databases, and comes with a data dictionary and a collection of scripts designed for data processing and analysis.

3 Process

Flowchart

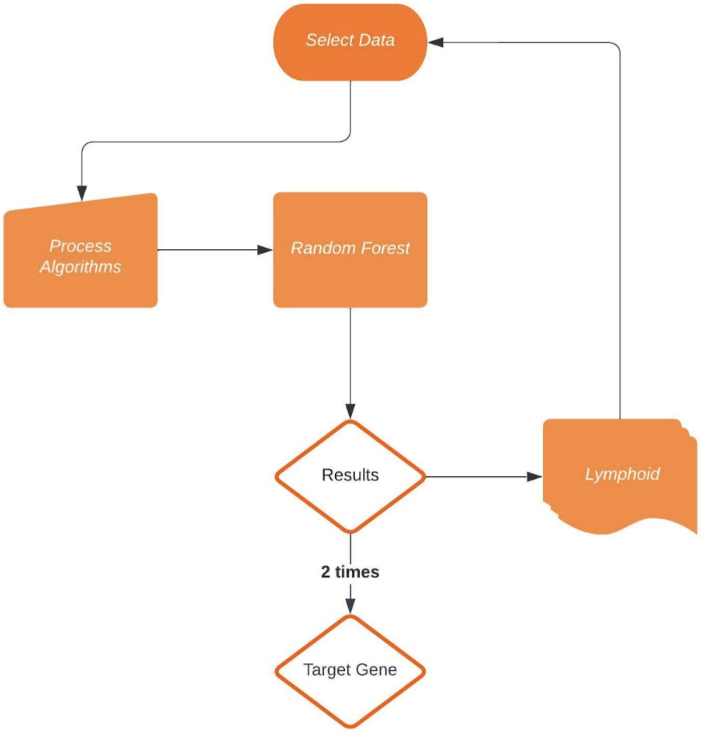


Figure 1: Page1 flowchart

3.1 Select Data

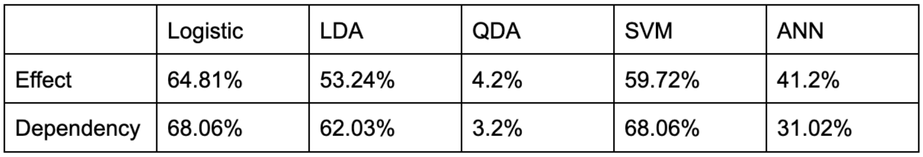
We selected two files, CRISPRGeneEffect and CRISPRGeneDependency, as our database. CRISPRGeneEffect represents the impact of CRISPR-mediated gene editing on cancer cell viability, while CRISPRGeneDependency denotes the degree to which a cancer cell line depends on a specific gene for its survival or growth.

For model training, we utilize the Effect score and Dependency score. These scores are calculated as log2 fold changes by taking the logarithm of the sgRNA counts for a gene at each time point, divided by the average sgRNA counts of that gene at the control time point. The numbers in the CSV data are transformed into the log-norm file using the following formula: Log2(Reads from an individual perturbation ÷ Total reads in PCR well * 106 + 1), which represents gene effect and dependency.

The Model file confirms that target genes originate from 20 distinct lineages.

4 Basic prediction

Before initiating our process, we plan to use various models to predict the results and accuracy scores for both datasets. In this step, we apply all the methods we have learned this semester, obtaining five different results from five distinct methods. Lasso regression was also attempted, but the results were suboptimal due to uncertainty in the alpha value, leading us to discard Lasso regression.



	Logistic	LDA	QDA	SVM	ANN
Effect	64.81%	53.24%	4.2%	59.72%	41.2%
Dependency	68.06%	62.03%	3.2%	68.06%	31.02%

Figure 2: Basic prediction

From the above analysis, the QDA score is unsatisfactory, primarily because the dataset contains numerous features, causing us to abandon this method. Moreover, the ANN score performs worse than the other three methods for prediction, likely because we lack a well-trained model for ANN. As observed, the remaining three methods demonstrate satisfactory performance for these datasets. Ultimately, we decide to eliminate the QDA method and continue utilizing the other four methods.

5 Random Forest

In this step, we plan to apply the random forest for feature selection in both datasets and subsequently perform the same processes as before to evaluate whether the random forest improves the accuracy scores.

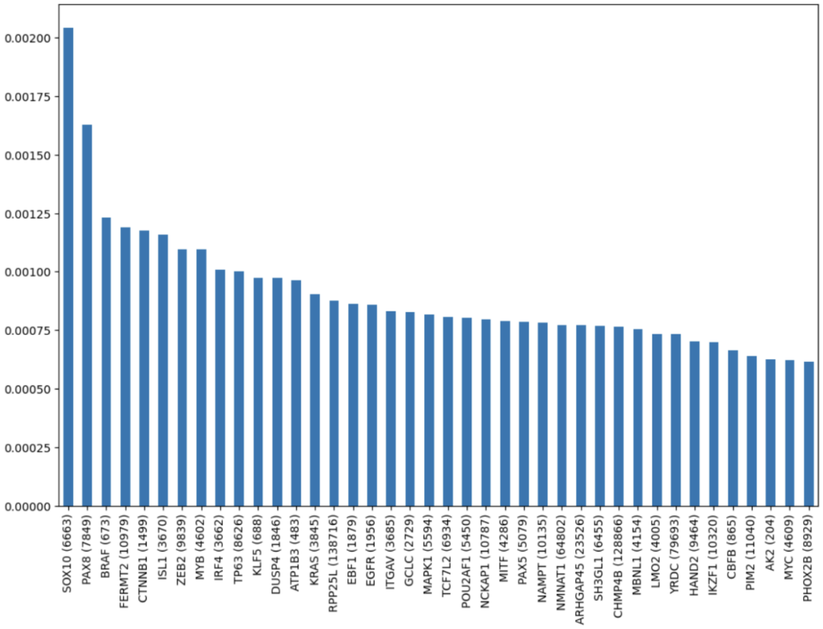
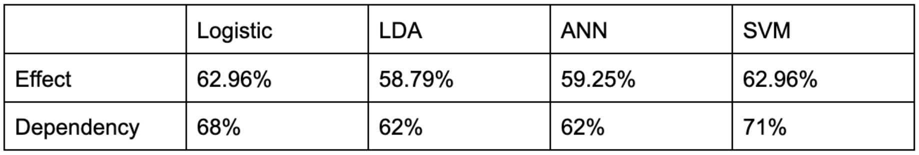


Figure 3: Datasets

Following random forest-based feature selection for the GeneEffect dataset, we obtain the results, and the same process is conducted for the GeneDependency dataset. The random forest ranks all features in the dataset; however, the image provided only presents the top 40 features.

6 2000 Features Classification

After the random forest feature selection, we repeat the same process as our basic regression to assess the improvement of our models following the random forest. This time, we only use 2000 features to train the models.



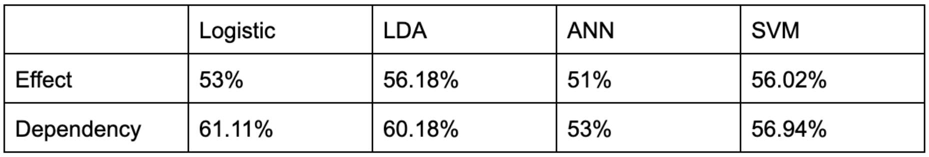
	Logistic	LDA	ANN	SVM
Effect	62.96%	58.79%	59.25%	62.96%
Dependency	68%	62%	62%	71%

Figure 4: 2000 Features

These results indicate that, after basic regression, the ANN model exhibits significant improvement after feature selection, likely due to the reduction in the number of features. Moreover, the other three methods display slight changes as the number of features decreases, demonstrating that our process is useful and effective.

7 2000 Features Classification

To enable a more in-depth comparison, we decide to further reduce the number of features for a more apparent contrast.



	Logistic	LDA	ANN	SVM
Effect	53%	56.18%	51%	56.02%
Dependency	61.11%	60.18%	53%	56.94%

Figure 5: 200 Features

In this iteration, we use only 200 features to train the model. The results indicate that all four methods yield fairly good outcomes even with just 200 features, further demonstrating the usefulness and feasibility of our process. After random forest feature selection, we may have successfully identified genes related to cancer.

8 Selecting a Specific Cancer Type for Prediction

From our analysis, we identify two cancer types, lymphoid and soft tissue, that exhibit relatively high performance in our model.

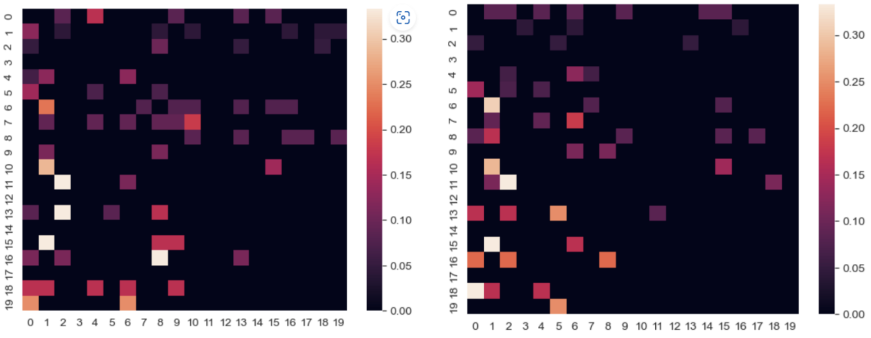


Figure 6: Heat Map

The heat map reveals that cancer number 3 and cancer number 14, which represent lymphoid and soft tissue cancers, respectively, have a high degree of overlap between the test data and predicted data. This suggests that our model achieves higher accuracy in predicting lymphoid and soft tissue cancers. However, we must exclude soft tissue cancer from further analysis, as the number of patients with this cancer type is too low to draw reliable conclusions.

9 Second Random Forest and Model Application

For this iteration of the random forest, we exclude all data except for lymphoid data.

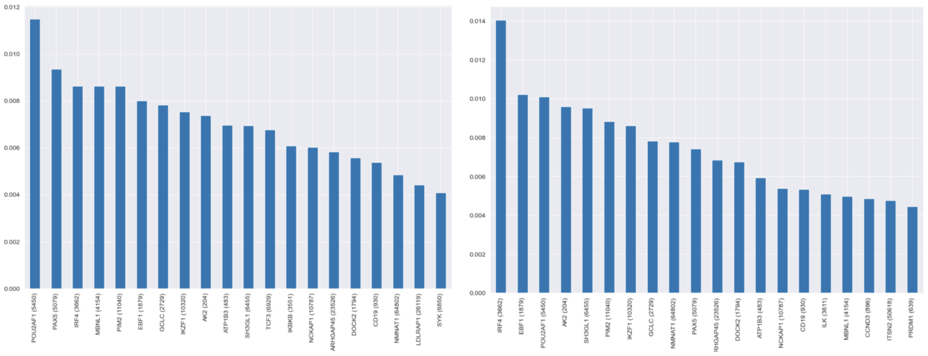


Figure 7: the Top 20 Features

After feature selection, we obtain the top 20 features. We then apply the same models as before, but focus on SVM and Logistic Regression, as these two models demonstrated better performance in our previous tests.

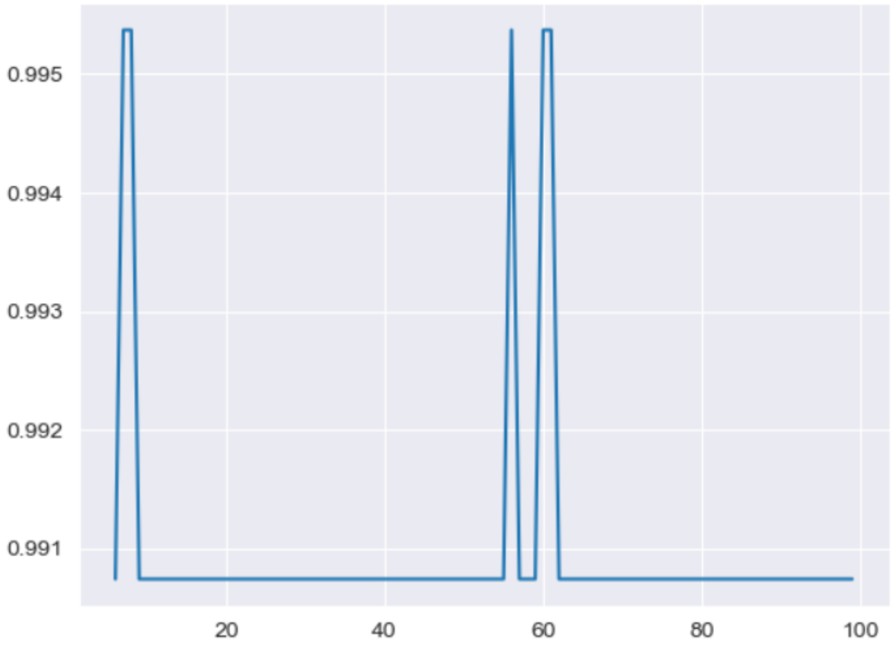


Figure 8: the Prediction Accuracy Scores

The graphs represent the prediction accuracy scores from the logistic models for one dataset. We use a for loop to test the accuracy as the feature size increases, and we obtain promising results for prediction. We observe similar results in another dataset and with SVM prediction. As seen, we achieve higher performance when the feature size is less than 20. This suggests that these top 20 genes may be related to lymphoid cancer.

10 Actual accuracy

Although we have achieved higher accuracy scores for prediction, it is essential to compare the actual accuracy to assess the real-world applicability of our prediction approach. To do this, we calculate the actual accuracy as the number of actual lymphoid cancer cases divided by the number of predicted lymphoid cancer cases.

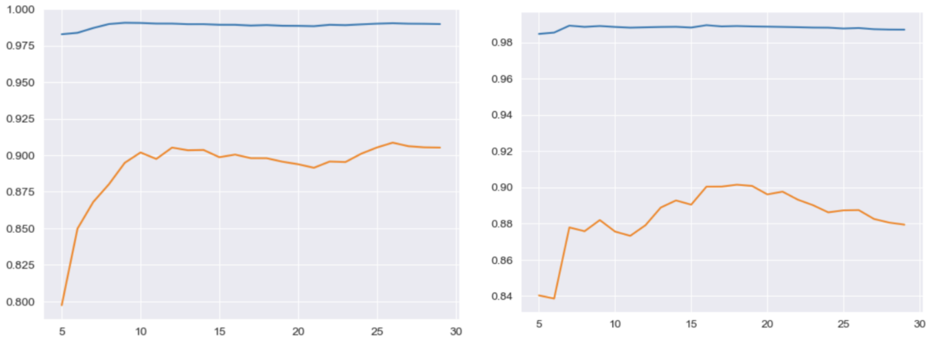
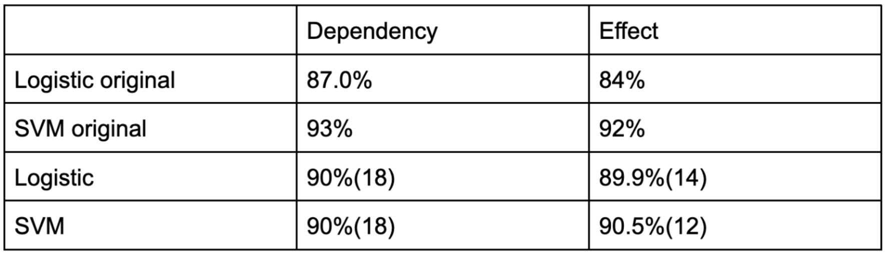


Figure 9: SVM Results from Two Datasets

The graphs presented show the SVM results from two datasets, with similar outcomes observed for logistic methods. The blue line represents our prediction, while the orange line represents the actual data. As we can see, when focusing on the top 20 features, we achieve the most accurate prediction scores, closely approximating the actual data.

11 Comparing Results and Conclusion

To determine the optimal feature size for the most accurate prediction scores, we plan to construct a for loop that iterates through various feature sizes and compares the resulting prediction scores to the actual accuracy. This process will help us identify the best-performing model and feature size for predicting lymphoid cancer, ultimately enabling more effective real-world applications of our approach.



	Dependency	Effect
Logistic original	87.0%	84%
SVM original	93%	92%
Logistic	90%(18)	89.9%(14)
SVM	90%(18)	90.5%(12)

Figure 10: Best Accuracy Score in the Selected Feature Size

In conclusion, the random forest data selection process has significantly improved the prediction accuracy of the models. By selecting only the most important features (less than 20 features) compared to the original dataset with more than 17,000 features, the machine learning models were able to achieve better actual accuracy scores. This demonstrates the efficiency and effectiveness of the random forest data selection method in refining the dataset and reducing the dimensionality of the data.

Comparing the accuracy scores from the original dataset (before random forest data selection) to those after the random forest data selection, it is evident that the latter led to more accurate predictions. This means that the random forest data selection method helped identify the most relevant features for predicting cancer cell lineages, thus significantly improving the performance of the machine learning models.

Overall, the use of random forest data selection in combination with machine learning models has led to better understanding and classification of cancer cell lineages. This approach has the potential to greatly impact cancer research and treatment by identifying relevant genes and providing more accurate predictions.

12 Limitation

There are several limitations in our results that may impact their accuracy:

12.1 Random Forest:

Firstly, we only use the Random Forest algorithm for feature selection. Another similar algorithm, Extra Trees (Extremely Randomized Trees), could be considered for better performance. Combining the results of these two algorithms in an ensemble learning approach may improve the accuracy and robustness of the model.

12.2 Accuracy Score:

Our process lacks an accuracy score component during the random forest phase. The accuracy score is crucial for evaluating the performance of a random forest model, as it measures the percentage of correct predictions made by the model. Incorporating an accuracy score could provide insights into the model’s ability to classify new data accurately.

12.3 Database:

Our study focuses on cancer genes as the experimental group but lacks a healthy gene pool as a control group. Additionally, the lymphoid data is still limited in size. These factors may affect the reliability and accuracy of our study since having a control group helps determine if the observed results in the experimental group are related to differences in the healthy genome. In the absence of a healthy gene pool and a sufficiently large experimental group, researchers may need to explore other control methods to reduce experimental errors.

13 Conclusion

Based on the results, it is possible to extract 20 significant genes from the CRISPRGeneEffect and CRISPRGeneDependency files after processing. The common significant genes are as follows:

Effect:[’POU2AF1 (5450)’ ’PAX5 (5079)’ ’IRF4 (3662)’ ’MBNL1 (4154)’ ’PIM2 (11040)’ ’EBF1 (1879)’ ’GCLC (2729)’ ’IKZF1 (10320)’ ’AK2 (204)’ ’ATP1B3 (483)’ ’SH3GL1 (6455)’ ’TCF3 (6929)’ ’IKBKB (3551)’ ’NCKAP1 (10787)’ ’ARHGAP45 (23526)’ ’DOCK2 (1794)’ ’CD19 (930)’ ’NMNAT1 (64802)’ ’LDLRAP1 (26119)’ ’SYK (6850)’]

Dependency:[’IRF4 (3662)’ ’EBF1 (1879)’ ’POU2AF1 (5450)’ ’AK2 (204)’ ’SH3GL1 (6455)’ ’PIM2 (11040)’ ’IKZF1 (10320)’ ’GCLC (2729)’ ’NMNAT1 (64802)’ ’PAX5 (5079)’ ’ARHGAP45 (23526)’ ’DOCK2 (1794)’ ’ATP1B3 (483)’ ’NCKAP1 (10787)’ ’CD19 (930)’ ’ILK (3611)’ ’MBNL1 (4154)’ ’CCND3 (896)’ ’ITSN2 (50618)’ ’PRDM1 (639)’]

Common: POU2AF1 (5450)’ ’PAX5 (5079)’ ’IRF4 (3662)’ ’MBNL1 (4154)’ ’PIM2 (11040)’ ’EBF1 (1879)’ ’GCLC (2729)’ ’IKZF1 (10320)’ ’AK2 (204)’ ’ATP1B3 (483)’ ’SH3GL1 (6455)”NCKAP1 (10787)’ ’ARHGAP45 (23526)’ ’DOCK2 (1794)’ ’CD19 (930)’ ’NMNAT1 (64802)’

Utilizing machine learning for cancer cell lineage classification offers several advantages over traditional methods, including:

1.Increased accuracy: Machine learning algorithms can analyze large amounts of data and identify patterns that may be difficult for humans to detect, leading to more accurate classification of cancer cell lineages.

2.Faster analysis: Machine learning algorithms can process data quickly, allowing for faster analysis and classification of cancer cell lineages.

3.Improved efficiency: By automating the classification process, machine learning algorithms can reduce the workload of researchers and healthcare professionals, freeing up time and resources for other important tasks.

4.Predictive modeling: Machine learning algorithms can be used to develop predictive models that can identify high-risk cancer cell lineages and help guide treatment decisions.

5.Integration with other data sources: Machine learning algorithms can integrate data from multiple sources, such as genomics, proteomics, and clinical data, providing a more comprehensive understanding of cancer cell lineages and improving treatment outcomes. In conclusion, using machine learning classification of cancer cell lineages can lead to more accurate and efficient analysis, improved treatment outcomes, and a better understanding of the underlying mechanisms of cancer development and progression.

References

[1] Viswanatha, Raghuvir, et al. “Figures and Data in Pooled Genome-Wide CRISPR Screening for Basal and Context-Specific Fitness Gene Essentiality in Drosophila Cells.” ELife, ELife Sciences Publications, Ltd, 27 July 2018.

[2][Murphy] Machine Learning: A Probabilistic Perspective by Kevin P. Murphy (Good details but very long book)

[3][SSBD] Understanding Machine Learning: From Theory to Algorithms by Shai Shalev-Shwartz, Shai Ben-David